Modeling language learning using specialized Elo ratings

Hou, Jue

The Association for Computational Linguistics
2019-08

Hou , J , Koppatz , M , Hoya Quecedo , J M , Stoyanova , N , Kopotev , M & Yangarber , R 2019 , Modeling language learning using specialized Elo ratings . in H Yannakoudakis , E Kochmar , C Leacock , N Madnani , I Pilán & T Zesch (eds) , Innovative Use of NLP for Building Educational Applications : Proceedings of the 14th Workshop . The Association for Computational Linguistics , Stroudsburg , pp. 494-506 , Workshop on Innovative Use of NLP for Building Educational Applications , Italy , 02/08/2019 . < https://www.aclweb.org/anthology/W19-4451 >

http://hdl.handle.net/10138/304628

Modeling language learning using specialized Elo ratings

[†]**Jue Hou**, [†]**Maximilian W. Koppatz**, [†]**José María Hoya Quecedo**,
[⋆]**Nataliya Stoyanova**, [‡]**Mikhail Kopotev** and [†]**Roman Yangarber**
[†]University of Helsinki, Department of Computer Science, Finland
[‡]University of Helsinki, Department of Modern Languages, Finland
first.last@helsinki.fi
[⋆]University of Milan, Department of Foreign Languages and Literatures, Italy
first.last@unimi.it

Abstract

Automatic assessment of the proficiency levels of the learner is a critical part of Intelligent Tutoring Systems. We present methods for assessment in the context of language learning. We use a specialized Elo formula used in conjunction with educational data mining. We simultaneously obtain ratings for the proficiency of the learners and for the difficulty of the linguistic concepts that the learners are trying to master. From the same data we also learn a graph structure representing a domain model capturing the relations among the concepts. This application of Elo provides ratings for learners and concepts which correlate well with subjective proficiency levels of the learners and difficulty levels of the concepts.

1 Introduction

A key goal of Intelligent Tutoring Systems (ITS) is to provide students with individualized learning, and thus support their learning process. Several systems for various subjects have proven to be effective, (Ritter et al., 2007; Arroyo et al., 2014; Klinkenberg et al., 2011). Our work is part of an international collaborative effort to developing large-scale Intelligent Computer-Aided Language Learning (ICALL) systems for use in real-world environments. Our system, Revita (Katinskaia et al., 2018), is in use in official university-level curricula at several major universities, to enhance language learning and teaching.[1]

To develop automated methods for personalized tutoring, robust models for assessing the current proficiency of learners are required. Such models enable quantitative evaluation and comparison of teaching methods with respect to the rate of development of the learners, which enables us to evaluate the performance of the ICALL system.

Several approaches for modeling knowledge have been proposed, such as Bayesian Knowledge Tracing (Corbett and Anderson, 1994), Learning Factor Analysis (Cen et al., 2006), and its more advanced variant—Performance Factor Analysis, (Pavlik Jr et al., 2009). These models, however, are complex and time-consuming to implement for ITS, and require large amounts of data for each learner, (Pelánek, 2016).

In this work, we present a simple and effective method for assessing the proficiency of language learners, as well as the difficulty of linguistic concepts, by utilizing the Elo formula, (Elo, 1978)—in an unsupervised fashion. The result is a set of rated concepts and a method for assessing the current and historical proficiency of each learner. We also present a concept graph—learned from the educational data—representing the dependencies between concepts.

We use educational data, collected from real language learners, in two assessment contexts. One context is where learners take language proficiency tests. During the test the system samples questions from a database, each question linked to a specific linguistic concept. Examples of concepts are collocations, certain type of verb inflection, certain word-order rules, etc. In the second context, learners do exercises based on "authentic" texts—chosen based on the learners' interests—that have a difficulty estimate, assigned by a statistical model. In each context, the result is a numeric Elo-based rating describing learners' proficiency in the target language.

The paper is organized as follows. Section 2 outlines out data generation process. In Section 3, we describe the Elo rating system, and discuss our modifications to the formulas in the two assessment contexts. In Section 4, we discuss our approach for building a domain model, represented by a concept graph. Section 5 describes simula-

[1] https://revita.cs.helsinki.fi/

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 494–506
Florence, Italy, August 2, 2019.

tions and experiments on obtaining Elo ratings for concepts and for assessing learner competency. In Section 6, we discuss preliminary results on correlation between Elo ratings in different assessment contexts, and the correlation between the Elo ratings and the performance levels assigned by teachers. We also demonstrate a graph of concepts, each with Elo difficulty estimates. In Section 7, we discuss current problems. Section 8 concludes with current directions of research.

2 Data

This work builds on educational data we have collected through a collaborative effort with language teachers at several universities. In this paper, we focus on students learning Russian as a second language (L2) at different levels on the CEFR scale, ranging from A1 to C2, (Little, 2007). The students yield data in two assessment contexts:

Language tests: In the testing context, the students take online language tests on a platform provided by the system. Each test is time-limited, and comprised of approximately 300 test items, sampled from a database of 3390 multiple-choice questions. The questions were prepared by language teachers and linguistic experts over a period of 20 years, (Kopotev, 2012, 2010). Each question is linked to one of 140 linguistic "concepts," also predefined by the experts. At the time of this writing, the response data consists of 600 000 test answers, by approximately 1000 learners. For each answer we record to which concept the question belongs, whether the answer was correct, as well as timestamps.

Language practice: In the practice context, students practice the language by doing exercises based on a text. The response data consists of student answers to the exercises: one set of exercises is associated with a *snippet* of text (e.g, one paragraph). The system offers various types of exercises, including multiple-choice questions, "cloze" quizzes (fill-in-the-blank), and listening comprehension, which are generated automatically based on the text chosen by the learner. Each text has been assigned a difficulty rating by a linear model, described in 3.4. Currently, we do not use information about to which linguistic concepts each exercise belong. For each attempted set of exercises, we use the percentage of correct answers, text difficulty, and the timestamp.

3 Rating methods

This section describes how we calculate and update ratings of users and linguistic concepts. We cover the two assessment contexts in which the data were generated: first, the testing context, and second, the practice context. The estimator of text difficulty is described at the end of this section.

3.1 Elo ratings

The Elo rating, introduced by Arpad Elo, (Elo, 1978), was originally used for rating the skills of chess players, and evolved versions of it are now widely used in a variety of domains, ranging from video-games to Tinder.

The Elo formula defines the *expected* result of actor A in a match against actor B according to:

$$E_A = \frac{1}{1 + 10^{\frac{R_B - R_A}{\sigma}}} \quad (1)$$

E_A is a value between 0 and 1, indicating the expectation (probability) of success/win. R_X refers to the current rating of actor X, and σ is a scale variable. The scale is traditionally set to 400 (in chess). This controls the spread of the resulting ratings. In the present work, we used $\sigma = 600$.[2]

The rating of actor A *after* a match with another actor is completed is updated according to:

$$R'_A = R_A + K(S_A - E_A) \quad (2)$$

The factor K controls the maximum rating adjustment that is possible at one time. We use a static K value of 32. S_A denotes the outcome, or the score of a match for actor A: loss, draw and win for A is denoted as 0, 0.5 and 1, respectively.

The Elo rating method has three important properties. First, the formulas are symmetric: A and B switch places when calculating with respect to B. This leads to the zero-sum property of the Elo rating distribution, when K is static—the amount of Elo lost by the losing actor is transferred to the winning one. As a result, the mean of the ratings of all actors will be whatever the initial rating is set to. We initially set the rating of all players (concepts and learners) to 1500.

Second, the magnitude of the rating update depends on the difference between the outcome S_A

[2]The reason for this is largely aesthetic; the number of users is currently small (about 1000), which results in a somewhat narrow spread of the rating distribution. A larger spread is more comparable to actual chess ratings, which is more familiar and easier to conceptualize.

and the expected score E_A. This means that a highly rated actor failing against a significantly lower rated one will have a severe loss in Elo, approaching the value of K. Conversely, the success of a higher rated actor is expected, thus it yields a small update in the rating. If the actors are evenly matched the update is between the two extremes.

Finally, the final, or "*current*," distribution of Elo ratings in a system *depends on the order* of the matches. This implies that the Elo rating is a representation of an actor's *current* proficiency. Consider an actor who fails in the first half of 100 matches, and succeeds in the second half, and compare the result to the reverse order. Because the rating updates decrease toward the expected extremes, and are exaggerated at the unexpected extremes, the resulting rating of the actor will be high when his successes are in the second half of the matches and low in the reverse case—even though the two sequences are events are the same, and differ only in their ordering.

3.2 Language test ratings

In (Klinkenberg et al., 2011; Pelánek, 2016), it was shown that Elo ratings can be adapted for use in educational systems, to model the proficiency of students and the difficulty of questions. In our system, in the testing context, the analogue of an outcome of a "game" is a student attempting an exercise; the two rated "actors" are the student and the exercise. S_A represents whether a student has answered the question correctly. The rating R_Q of a question Q captures the difficulty of the question.

An adjustment is made for *multiple-choice* problems to account for the fact that students have some chance of guessing correctly, even if they do not know the correct answer. For this, we adopt an approach recommended by Pelánek (2016), penalizing the expected value by the probability that a random guess is correct. A similar expectation formula is proposed in Item Response Theory (Embretson and Reise, 2013).

In (Klinkenberg et al., 2011; Pelánek, 2016), the focus was assessing the difficulty of each particular *question*. In our system, we model the concept category to which the questions belong; every question from a concept is a representative for the concept in the calculations. In our data, each question is linked to one linguistic concept. The formula for the expectation of student S having a correct response on a question from Concept C adjusted for guessing is:

$$E_A = \frac{1}{k} \cdot 1 + (1 - \frac{1}{k}) \cdot \frac{1}{1 + 10^{\frac{R_C - R_S}{\sigma}}}, \quad (3)$$

where k is the number of choices in the multiple-choice question. We expect the Elo ratings for concepts to approach their true value after a large number of data points ("games" or exercise attempts) have been sampled. To obtain concept ratings of better quality, they are learned by re-adjusting all ratings by re-playing all games/attempts in chronological order over several epochs. We call this the Elo "*burn-in*" period, and describe it in Section 5.1. This is used to obtain "stable" concept ratings.

3.3 Language practice ratings

In the practice context, the exercises are different from the test context. Our method therefore differs in the two contexts in several ways. Here, the exercises sets that the learners have answered are composed of a variety of exercises, some of which may not be linked to a specific linguistic concept (recall, each test item is linked to a concept). This means we cannot share the ratings directly between the two learning contexts.

To address this problem, we make the simplifying assumption that *on average* the exercises in a given text correspond to the *difficulty* rating of the text. The method for estimating the difficulty of a text is described in the following subsection.[3]

Concretely, we define S_A as the percentage of correct responses in a given set of exercises. E_A for a set of exercises is set to the Elo rating of the *entire text* from which the exercises are drawn.

We update the learner's Elo after *each* set of exercises, but update the text's Elo only after an entire iteration through the text. Crucially, the system updates the Elo for the text only with respect to the *specific user* who practiced it. This models the notion that as I practice with the same text over and over, the text becomes "easier" for me—but not for other players.

3.4 Estimator of text difficulty

Modeling and characterizing the readability of texts is a well-studied problem with a long history, (Dubay, 2009). Experiments with lexical and

[3]This means that A. our difficulty model should return an accurate estimate of the complexity of the text, and B. that these difficulty estimates should be properly calibrated to the desired rating scale.

grammatical features have been conducted, (Chen and Meurers, 2016; Heilman et al., 2008).

We use a simple linear model for estimating the difficulty of a given text. The output of this model is scaled onto the Elo rating scale. This enables us to calculate the expected result of any rated learner solving exercises from any rated text.

Lexical frequency is known to be a powerful predictor of text readability, (Chen and Meurers, 2016). We use the normalized mean of the lexical frequencies of the tokens in a text as a feature. Additionally, we use **mean token length** and **mean sentence length**, as they are also used in classic readability measurements, (Kincaid et al., 1975; Flesch, 1979). These three features are scaled for a simple 3-variable regression model.

Currently, data for training this model are partitioned into two types: texts from sources with simplified language vs. texts from difficult sources. We label simple texts with a value of 0.2 and difficult texts with 0.8.[4] We aim for a model that produces a correct ordering of texts with respect to their difficulty. We do not need an exact estimate. If the learner Elo ratings based on these estimates correlate well with the ratings from the testing context, we consider this estimator accurate enough. In Section 5 we show that so far, this indeed seems to be the case.

The model outputs typically range between 0 and 1. We scale these values to Elo ratings according to formula 4. This transformation is based on the Elo rating distribution acquired in the testing context. The bounds are clamped at $(0, 1)$, as some texts may get a high difficulty value (sometimes even > 2, in extreme cases).

$$f(x) = \begin{cases} 600, & \text{if } x < -0.4 \\ 1000x + 1000, & \text{if } -0.4 \leq x \leq 1.4 \\ 2400, & \text{if } x > 1.4 \end{cases} \quad (4)$$

4 Concept graph

We are interested in finding a model for the "natural" order in which learners acquire linguistic concepts—directly from learner data. Learners will find some orders more natural than others—e.g., when some concept is a requirement for another. The fact that one concept "precedes" another is called the *surmise* relation in Knowledge Space Theory (Doignon and Falmagne, 1999).

Often the domain model in ITS is built by eliciting domain knowledge from experts. We devised a baseline model to infer such relations from user data, without supervision. As we mentioned, each test question is mapped to a certain linguistic concept, and we store all test results from all students. Based on these results, we know to what extent which users have mastered which concepts; from this, we can try to tell apart the more basic concepts from the more advanced ones.

The aim is to build a partial order over the set of all concepts C, which specifices which concepts are related—i.e., we write $c_2 \rightarrow c_1$ to mean concept c_2 *presupposes* (or implies) concept c_1.[5]

Given a set of users U, we build a matrix of "mastery" scores M, of dimension $|U| \times |C|$. Every element M_{ij} is the proportion of correct answers that user u_i has given for concept c_j. In the current implementation, we consider each "user" to be a single *test session*. If the same person completes the test at different times, they will be treated as different users for the purpose of computing consistency. This is done to take into account the fact that a user's level of proficiency changes over time.

For every pair $c_1, c_2 \in C$, we check whether $c_2 \rightarrow c_1$, $c_1 \rightarrow c_2$ or $c_1 \perp c_2$. We compare all columns of M pairwise. Let c_j denote column j of M and c_k column k. To check whether $c_k \rightarrow c_j$ is true, we define a logical function CON_u which checks that user u is consistent with this relation:

M_{uk}	M_{uj}	$CON_u(c_k \rightarrow c_j)$
0	0	1
1	0	0
0	1	1
1	1	1

Again, here c_k is the "harder" concept than c_j. Thus, if $M_{uk} = 0$ (user u knows nothing about concept c_k), that is consistent with $c_k \rightarrow c_j$ regardless of the value of c_j. Conversely if $M_{uj} = 1$ (user u mastered concept c_k perfectly), that is consistent with $c_k \rightarrow c_j$ regardless of the value of c_k.

In practice, the values in M are fractions between 0 and 1. Therefore we introduce two thresh-

[4] This is a simplification, which does not reflect reality accurately, as not all texts from a given source are of equal difficulty. However, simply fitting a low dimensional, high-bias estimator such as this yields a reasonable baseline model that generalizes well enough to other texts. We will explore more sophisticated models in the future.

[5] We would like to say that c_1 is a *prerequisite* for c_2, but that may be too strong a claim. However we *may* be able to learn from the data that *typically* c_1 is mastered before c_2.

old parameters to map M_{ij} to zeroes and ones: we believe that a user u really does not know the (harder) concept c_k if $M_{uk} \leqslant \overline{\theta}_{guess}$: the "guessing" upper bound, below which we believe that the user does not know the concept, while sometimes only guessing the correct answer. Analogously, we say that a user u has mastered quite well the (easier) concept c_j if $M_{uj} \geqslant \underline{\theta}_{master}$: the "mastery" lower bound, above which we believe that the user knows the concept, while sometimes making a few mistakes.

To check whether $c_k \rightarrow c_j$, we compute the proportion σ of all users who are consistent with this relation $c_k \rightarrow c_j$, as follows:

$$\sigma(c_k \rightarrow c_j) = \frac{1}{n} \sum_{u \in U} CON_u(M_{uk}, M_{uj})$$

where:

$$CON_u(M_{uk}, M_{uj}) = \begin{cases} 1 & \text{if } M_{uk} \leqslant \overline{\theta}_{guess} \\ 1 & \text{if } M_{uj} \geqslant \underline{\theta}_{master} \\ 0 & otherwise \end{cases}$$

n is the total number of users. θ_{guess} and θ_{master} refer to the thresholds of guessing and mastering respectively. That is, we ignore all users where the level of proficiency in c_1 and c_2 suggests that user only partially understands both concepts, since they do not support to the consistency from $c_k \rightarrow c_j$.

We then apply a *consistency* threshold θ such that, if $\sigma(c_2 \rightarrow c_1) > \theta$, we add the relation $c_2 \rightarrow c_1$ to our partial order.

Finally, we represent the partial order as a directed acyclic graph (DAG), where each path in the graph represents a possible prerequisite route toward learning a concept. For example, if we wish to obtain a complete syllabus for a language course, then we can find a total order compatible with our partial order (i.e., a linear extension) by topologically sorting the nodes in the graph.

We tested this approach with a set of over 620K answers gathered from 700 users, and manually evaluated the results, setting $\theta = .7$.

Our domain experts confirm that the resulting graph provides a plausible model for the relations between the concepts in the language.

5 Experiments

5.1 Elo burn-in

In conventional Elo rating systems, the ratings of both actors are updated after each match. Our goal is to learn a stable rating for each language concept, which we can keep unchanged as learners take further tests and improve. The rationale behind this is that the concepts don't become less difficult for everyone if a learner masters it.

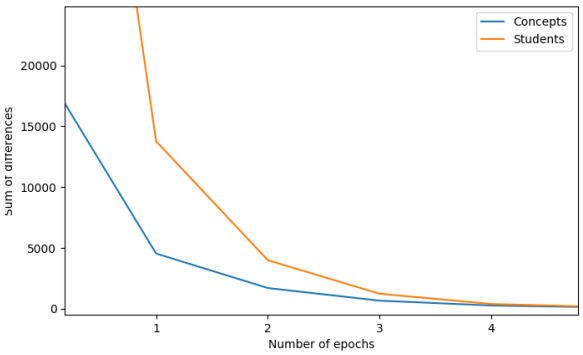

Figure 1: Learning concept Elo ratings with burn-in

To achieve this, we perform a burn-in simulation based on the user data we have collected. Specifically, we take the entire collection of data and calculate the Elo updates for each data point—user U doing exercise E—in chronological order. We do this for repeated epochs until the sum of differences between epochs nears zero. This is illustrated in Figure 1.

Once convergence is achieved, we reset the student ratings, and recalculate them from scratch with the *fixed* concept ratings. This yields a system where the learners' ratings are comparable and independent for any learner, including future ones.

5.2 Test and exercise simulation

To verify that one test (of 300 items) is sufficient for a learner to reach her current "true" rating, we performed simulation experiments. Concretely, we measured how many test items a new learner (e.g., with a rating of 1800 Elo) must answer in order to settle on her rating.

In this simulation, each student is initially rated as 1500, and complete some test items. We use the Elo expectation formula to get the response accuracy for the simulated actors. We perform randomized simulations of actors performing one 300-item test with this average response accuracy and observe the results, shown in Figure 2. The actors' responses—correct or false—are sampled randomly according to her supposed correct rate. We also add a small amount of normally distributed random noise. As Figure 2 shows, in this case, one 300-item test is enough to reach one's "true" rating.

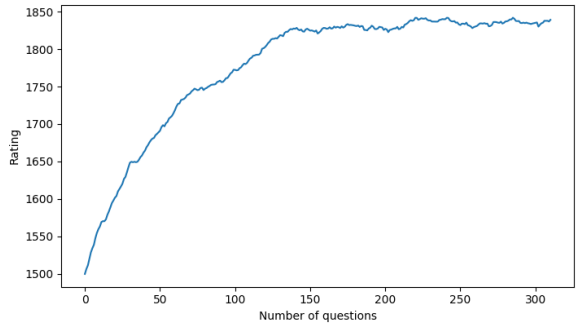

Figure 2: Simulation: test items

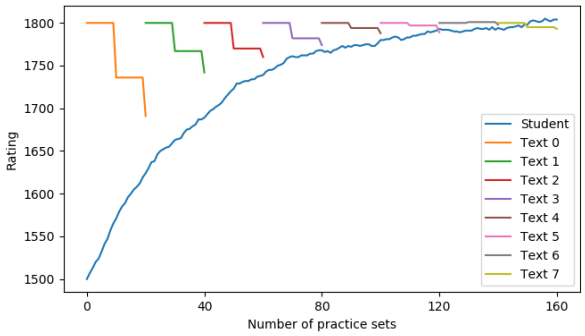

Figure 3: Simulation: exercises based on texts

In contrast to the test item context, in the context of text-based exercises, the Elo ratings will be updated over time both for for learners and practice texts. We conducted a simulation to visualize the rating trends.

In the exercise simulation, the simulated actor has the same starting point as in the simulation for tests—1500. We fix her rate of correct responses, e.g., as 50%, and the actor is practicing only simulated texts rated at 1800. This means that the theoretical upper bound of the actors' rating is, by definition, under 1800.[6] The same normally distributed noise is introduced in this simulation as above. The data in our system shows that texts typically contain 10 snippets (exercise sets) or fewer. After a full pass through a text, namely, 10 problem sets, the ratings for the text are updated with respect to this particular actor. In this simulation each text is to be practiced twice. Figure 3 shows the result of this simulation. We stopped the simulation after 8 texts (each practiced twice) because the rating of the actor had converged to 1800. Since the student begins far below 1800, the first several problem sets will have a substantial drop in their rating, due to the large initial difference between student's rating and text's difficulty estimate. This initial burn-in process soon ends, after the student practice several snippets. As we can observe in Figure 3, the initial burn-in process ends after 3 texts—10 snippets/practice sets per text, each text practiced twice, yielding 60 snippets—the subsequent ones showing only a slight change. The simulation is based on the assumption that the student will go though every problem set in strict order. In reality, this is not very likely. We can therefore infer that this student can reach a rating around 1800 rating after practicing with less than 140 snippets/problem sets.

[6] 50% success rate in a competition setting means the players are perfectly matched; since the text is rated 1800, the learner's Elo rating should also reach 1800 and remain there.

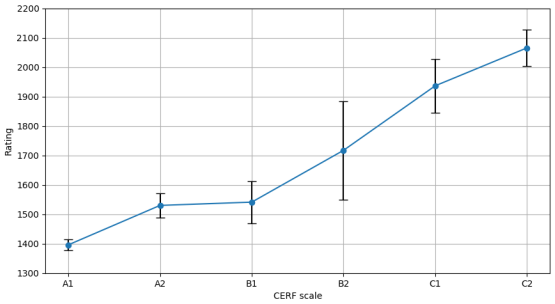

Figure 4: Expert annotated CEFR groups vs. mean Elo rating for each CEFR level

6 Results

The result of primary interest is how well the Elo ratings given by our system correspond to CEFR levels assigned to the learners by the expert teachers, who estimate the learners' proficiency based on a wide range of assessment criteria, including written essays and oral exams. The relationship between Elo and CEFR is illustrated in Figure 4.

The figure shows the means and 95% confidence intervals of the Elo ratings for the CEFR levels of students, assigned subjectively by the teachers. The numbers of students at each CEFR level in our experiments are given in Table 1. The data in Figure 4 comes from 142 students whose CEFR levels were established *independently* of these tests by the teachers. Although language competency encompasses several skills—reading, writing, aural comprehension, speaking—and the learner may be at different levels in different skills, we normally expect that the competencies across different sills are fairly well correlated. The correlation om Figure 4 is 0.90. Thus, the data in the figure indicate good correlation between our rating method and actual proficiency.

While assessing and modeling the improvement

CEFR	Students
A1	48
A2	31
B1	18
B2	6
C1	17
C2	22
Total	142

Table 1: Number of CEFR rated students in Figure 4

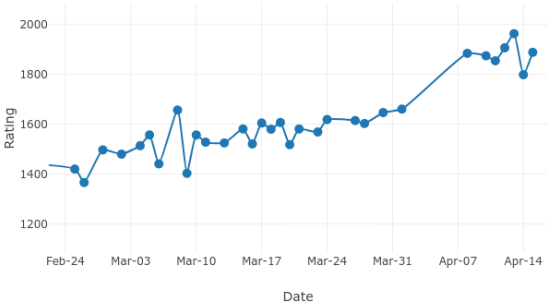

Figure 5: A representative student test Elo progress

of learners is largely in the realm of future work, we have interesting preliminary data of learners' improvement. Figure 5 shows an example of the progress of a typical actual learner taking tests. The learner takes (randomized) tests over a period of time, with the rating improving consistently. In this example, we see an increase of almost 300 in average Elo during the periods. The improvement is attributed to the fact that the learner has covered new material during the period, mastered more of the concepts in the test, and consequently scoring better on those questions.

Another crucial question is how well our Elo estimates from the exercise context correlate with Elo ratings from the testing context. For this we also have initial results, shown in Figure 6. We have so far collected only a modest amount of learner data, and therefore the conclusions drawn from the data are preliminary. Investigating this correlation requires substantial data from learners who have worked in both contexts: practicing with exercises based on texts, and completing tests. The figure shows results for the top 17 students, who have completed at least 1000 exercise sets (text snippets, in orange) and at least 300 test items (in blue)—sorted according to their *test* rating. The figure shows good correlation between the two rat-

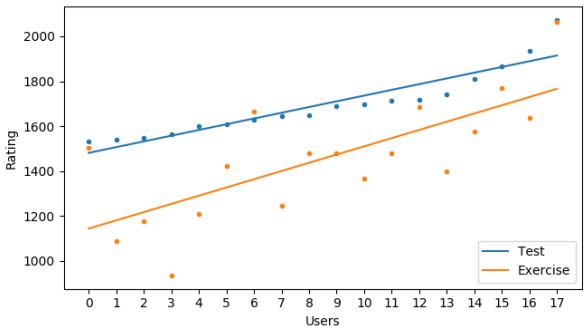

Figure 6: Correlation between test and text-based exercise ratings ($\rho = 0.79$)

ings for the students. At present, the correlation between test and practice Elo score is 0.79.

The last goal in our work is to investigate the relationship between the concept graph and the Elo ratings. Figure 7 shows a small sub-graph of the concept graph. The complete concept graph 8 can found in the Appendix. From Figures 7 and 8, we can see that typically (though not always), if concept A surmises concept B, A will have a higher Elo rating than B. This makes sense, as more difficult concepts should surmise easier ones. The Elo ratings and the graph structure do not correspond perfectly. Expecting that real-valued linear ratings can accurately describe the natural order of concepts is unreasonable. Since the graph and the Elo ratings describe different processes, it is not surprising to find inversions in the graph, such as between concepts 98 and 93, in Figure 7. The structure of the graph, its relation to the difficulty of concepts, and the natural learning order of concepts is an key future research topic.

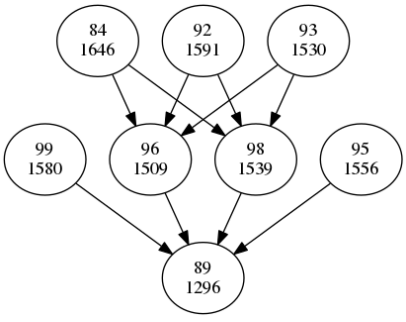

Figure 7: Sub-graph of the complete concept graph (see Supplementary Materials). Arrows denote implication. Top number: ID of the concept (appendix A); bottom number: Elo score of the concept (appendix B).

7 Current problems

We are in the process of collecting user data and evaluating the methods of assessment presented in this paper. Subjective CEFR ratings are being collected from the teachers for some of the students. While the number of hand-labeled CEFR ratings is modest at the time of writing, it is sufficient to indicate that using Elo-based ratings for measuring language proficiency shows promise. Building domain models based on the concept precedence graph is another direction of research.

Comparing our ratings and expert-annotated proficiency levels in larger quantities will raise the confidence of our method. We must note that a single value cannot be expected to describe the language proficiency of a learner completely, as there are several aspects of the language to master.

We plan to compare other Elo-based models, such as Glicko (Glickman, 1999), and TrueSkill (Herbrich et al., 2007), with our Elo rating formula. Robust methods for obtaining numerical estimates of skills enables us to develop ICALL systems, by facilitating the quantitative evaluation of skills, and the resulting improvement of the learners.

8 Conclusions

The main contributions of this paper are:

- We adapt and evaluate Elo-based rating formulas for modeling language learners' proficiency, as well as the difficulty of texts and linguistic concepts, not only the difficulty of questions/test items.
- We obtain static difficulty ratings for the linguistic concepts by performing an initial burn-in for the Elo ratings based on a large amount of learner data, and then assess students' proficiency using the learned Elo ratings. Feedback from language teachers/experts indicates that the ratings correlate with their estimates of learner competency.
- We use a linear-regression model of text difficulty as an estimator, to obtain Elo ratings for the texts. This enables us to rate the performance of learners, who practice with exercises generated from the texts. Preliminary results indicate that there is a positive correlation between ratings in the exercise and the test contexts.
- We build a partial order over all concepts found in the domain, and visualize the partial order as a DAG over concepts. The concept graph is not linear as the Elo ratings. The structure of the concept graph and the Elo ratings of the concepts generally agree in that the graph displays a strong tendency of decreasing rating from the more complex concepts to the more fundamental concepts, as indicated by the actual data collected from the learning process.

In sum, the proposed methods enable us to rate the proficiency of the current and future language learners, which is a fundamental goal in ITS and ICALL.

Acknowledgements

Work was supported in part by the Academy of Finland, Project *FinUgRevita*, Grant No. 267097, and by HIIT—Helsinki Institute for Information Technology.

References

Ivon Arroyo, Beverly Park Woolf, Winslow Burelson, Kasia Muldner, Dovan Rai, and Minghui Tai. 2014. A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 24(4):387–426.

Hao Cen, Kenneth Koedinger, and Brian Junker. 2006. Learning factors analysis – a general method for cognitive model evaluation and improvement. In *Intelligent Tutoring Systems*, pages 164–175, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 84–94.

Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278.

Jean-Paul Doignon and Jean-Claude Falmagne. 1999. *Knowledge Spaces*.

William Dubay. 2009. Unlocking language: The classic readability studies. *Professional Communication, IEEE Transactions on*, 51:416 – 417.

Arpad E. Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub., New York.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Rudolf Flesch. 1979. How to write plain English: Let's start with the formula. *University of Canterbury*.

Mark E Glickman. 1999. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pages 71–79. Association for Computational Linguistics.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill™: A Bayesian skill rating system. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel.

Sharon Klinkenberg, Marthe Straatemeier, and Han LJ van der Maas. 2011. Computer adaptive practice of maths ability using a new item response model for on-the-fly ability and difficulty estimation. *Computers & Education*, 57(2):1813–1824.

Mikhail Kopotev. 2010. Система прогрессивного тестирования Karttu (описание и первые результаты). Русский язык за рубежом, (3):23–29.

Mikhail Kopotev. 2012. Karttu: результаты языкового тестирования в школе и вузе. Формирование и оценка коммуникативной компетенции билингвов в процессе двуязычного образования: Сб. ст./Под ред. ЕЕ Юркова и др. СПб.: МИРС, pages 312–339.

David Little. 2007. The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4):645–655.

Philip I Pavlik Jr, Hao Cen, and Kenneth R Koedinger. 2009. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*.

Radek Pelánek. 2016. Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, 98:169–179.

Steven Ritter, John R. Anderson, Kenneth R. Koedinger, and Albert Corbett. 2007. Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, 14(2):249–255.

A List of concepts

ID	Rating	Name of concept
6	1602	Lexicology. Lexical semantics
7	1860	Collocations
8	1679	Lexicology. Coordination of words
9	1382	Verb. Case government
10	1541	Verb. Prepositional government
11	1550	Adjectives
12	1509	Nouns
19	1471	Impersonal verbs (except verba meteorologica) and their government
20	1482	Predicative adverbs and their government. Existence, state, time
21	1480	Predicative adverbs and their government. Necessity, possibility, impossibility
22	1577	Negative constructions with predicative 'be' (and synonyms), genitive of negation
23	1423	Sentences with dative subject: 'Кате 25 лет'
24	1445	Expressions of time, place and manner. Preposition-free expressions
25	1621	Constructions with cardinal numerals
26	1470	Constructions with collective numerals
27	1652	Genitive plural of Pluralia tantum words
28	1630	I declension. Type 'музей-музеи, воробей-воробьи'
29	1595	I declension. Type 'санаторий'
30	1662	I declension. Fleeting vowels and alternations я, е, ё / й ('заяц-зайца, заём-займа')
31	1608	I declension. Type 'карандаш'
32	1671	I declension. Type 'адрес-адреса'
33	1562	I declension. Type 'солдат-много солдат, сапог-пара сапог'
34	1586	I declension. Type '-анин/-янин, -ин'
35	1518	I declension. Type 'дерево-деревья'
36	1616	II declension. Type 'армия'
37	1603	II declension. Type на -ня
38	1713	II declension. Type 'статья'
39	1616	Fleeting vowel in genitive plural
40	1504	Nouns with prepositions в/на ending in -у/-ю in prepositional singular
41	1619	Nouns ending in -у/-ю in prepositional singular and -а in nominative plural
42	1644	Possessive adjectives. Type 'лисий'
43	1601	Ordinal numbers. Type 'третий'
44	1612	Possessive adjectives. Type 'мамин'
45	1543	Cardinal numbers. 'Сто' vs. 'пятьсот, шестьсот, семьсот, девятьсот'
46	1608	Quantifiers. Collective quantifiers in oblique cases
47	1528	Quantifiers. Collective quantifiers 'оба, обе'
48	1594	I conjugation. Type 'плакать'
49	1511	I conjugation. Type 'рисовать'
51	1560	II conjugation. Type 'молчать'
52	1674	Preterite. Type 'исчезнуть'
54	1570	Regular verbs with vowel alternation
55	1512	Resultative
56	1553	Iterative / potential iterative / qualities
57	1507	Expression of duration - 'за какое время'
58	1416	Factual meaning of verbs
59	1817	Aspect, expression of action completed in the past
60	1648	Aspect, expression of capability/incapability. ('Тебе этого не понимать/понять!')
61	1509	Inception of action
62	1525	'Забыть, успеть, удаться' + infinitive
63	1539	'Уметь, нравиться, любить' etc. + infinitive
64	1651	'Пора, скорее' + infinitive
65	1553	'Нельзя, невозможно, не могу' + infinitive
66	1499	'Не' + infinitive
67	1501	Negative sentences
68	1411	Imperative
69	1559	Proximal future
70	1674	Impersonal sentences. Infinitival sentences. Subordinate sentences ('если / прежде чем' + infinitive)
71	946	Impersonal sentences. Infinitival sentences. Modal expressions
72	1580	Impersonal sentences. Infinitival sentences. Sentences, with negative pronouns and adverbs
73	1701	Generis-personal sentences
75	1516	Passive and its relation to indefinite-personal sentences (equivalent of Finnish impersonal passive)
76	1800	Stress: forms of verbal preterite
77	1706	Stress: short adjectives
78	1736	Stress: participles
79	1548	Nouns. Type A: stress on the stem
80	1719	Nouns. Type B, B1, B2: stress on the ending

81	1633	Nouns. Type C, C1: stress on the stem in singular, on the ending in plural
82	1614	Nouns. Type D, D1: stress on the ending in singular, on the stem in plural
83	1610	Nouns. Type ‘нож-ножом, сторож-сторожем’
84	1646	Declension of adjectives. Type ‘хорошего’
85	1295	Place of adverbs of time, place and manner
86	1693	Place of negation in the sentence
87	1785	Place of participles in the sentence
88	1618	Place of gerunds in the sentence
89	1296	Place of pronouns in a phrase
90	1589	Word order in sentences introducing direct quotations
91	1681	Second person singular imperative in conditionals (‘если’)
92	1591	Usage of ‘сам’ and ’один’
93	1530	Sentences of type ‘Знаю его как врача.’
94	1567	Sentences of type ‘Быть грозе’
95	1556	Sentences of type ‘Лодку унесло ветром’
96	1509	Genitive plural
97	1537	Frequent prefixed verbs of motion + prepositional constructions
98	1539	Animate noun object
99	1580	Unprefixed verbs of motion
100	1121	Unstressed fleeting vowels in roots and suffixes of nouns and adjectives
101	1391	Unstressed vowels in verbal forms
102	907	Unstressed vowels in roots
103	1452	Unstressed vowels in case endings
104	1179	Unstressed vowels in prefixes
105	370	Unstressed vowels in suffixes
106	1557	Unstressed vowels linking compounds
107	1224	Unstressed particles не and ни
108	357	Letter г in ending -ого (-его)
109	1443	Letter й
110	1241	Letter ч and ш before н and т
111	1427	Letters ъ and ь:
112	1500	Vowels in verbal endings
113	1451	Vowels in the infinitive (indefinite form) before -ть
114	1365	Vowels not after sibilants and ц
115	1231	Vowels after sibilants and ц
116	1091	Voiceless and voiced consonants
117	1378	Consonant clusters at the juncture of morphemes
118	1654	Double and single -н- in suffixes of adjectives and nouns
119	2025	Double and single -н- in suffixes of full and short forms of adjectives
120	1706	Double and single -н- in suffixes of past passive participles and corresponding adjectives
121	1414	Double and single -н- in words derived from adjectives and participles
122	1341	Double consonants in borrowed roots and suffixes
123	1529	Double consonants in native roots
124	1131	Double consonants at morpheme juncture
125	997	Silent consonants
126	1050	Peculiarities of spelling of certain roots
127	1273	Peculiarities of spelling of certain suffixes
128	1868	Joint vs. separate spelling of negation ‘не’: verb (+participles)
129	1320	Joint vs. separate spelling of negation ‘не’: pronoun
130	1296	Joint vs. separate spelling of negation ‘не’: adverb (+‘несколько’)
131	1475	Joint vs. separate spelling of negation ‘не’: adjectives (+full-short)
132	1490	Joint vs. separate spelling of negation ‘не’: noun
133	1563	Joint vs. hyphenated spelling: adjectives
134	1590	Joint vs. hyphenated spelling: numerals
135	1059	Joint vs. hyphenated spelling: pronouns
136	1312	Joint vs. hyphenated spelling: adverbs
137	1555	Joint vs. hyphenated spelling: common nouns
138	1484	Joint vs. hyphenated spelling: function words
139	1836	Joint vs. hyphenated spelling: proper names
140	1555	Capitalized vs. lowercase: astronomical/geographical names
141	1664	Capitalized vs. lowercase: posts, titles, awards
142	1738	Capitalized vs. lowercase: names of official organizations
143	1898	Capitalized vs. lowercase: names linked to religion, historical epochs and events
144	1718	Capitalized vs. lowercase: names of trademarks, documents, works of art
145	1309	Capitalized vs. lowercase: proper names of persons, animals
190	1527	Accusative/ergative subject + Impersonal verb: ‘Васю тошнит’
191	1467	Prep+genitive subject + Impersonal verb: ‘У меня болит голова/шумит в голове’
230	1496	Dative subject + adverb: ‘мне (стало) плохо/нужно/скучно’
231	1548	Dative subject + impersonal verb: ‘мне идет/надоело/везет/хватит’
232	1431	Dative subject + impersonal-reflexive verb: ‘мне нравится/кажется/пришлось’

B Concept graph

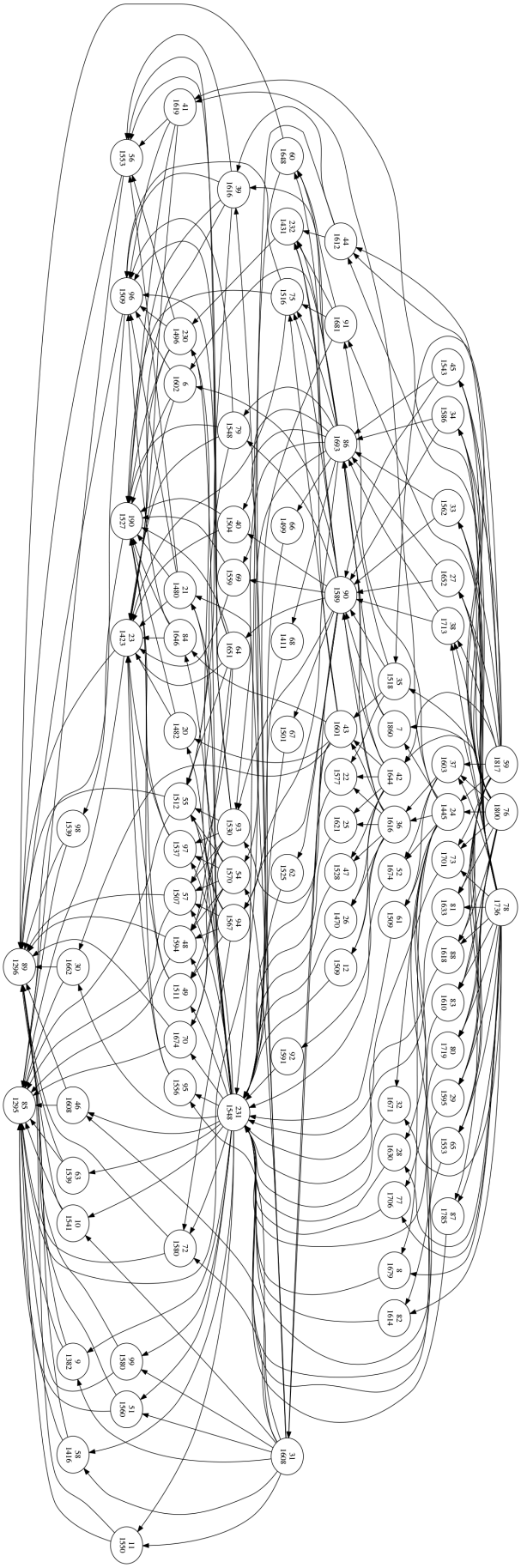

Figure 8: The full concept graph.